Human Genome Variation

DATA REPORT Open Access

Exploration of intermediate-sized INDELs by next-generation multigene panel testing in Han Chinese patients with breast cancer

Chihiro Hata[1], Hirofumi Nakaoka[1], Yu Xiang[2], Dong Wang[3], Anping Yang[4], Dahai Liu[4], Fang Liu[4], Qingfeng Zou[5], Ke Zheng[6], Ituro Inoue[1] and Hua You[3,5,7]

Abstract

Multigene panel testing via next-generation sequencing focuses on the detection of small-sized mutations, such as single nucleotide variants and short insertions and deletions (INDELs). However, intermediate-sized INDELs have not been fully explored due to technical difficulties. Here, we performed bioinformatics analyses to identify intermediate-sized INDELs in 54 cancer-related genes from 583 Han Chinese patients with breast cancer. We detected a novel deletion–insertion in a translational variant of *PTEN* (also known as *PTENα*) in one patient.

Breast cancer is the most common type of cancer among women[1], and approximately 10–15% of the cases are associated with hereditary mutations in DNA repair genes, including *BRCA1/2*[2]. With the advent of next-generation sequencing (NGS) technologies, genetic testing of *BRCA1/2* is now conducted worldwide. Multigene panel testing utilizing NGS technologies has enabled researchers to identify pathogenic mutations in genes other than *BRCA1/2*. It is also useful for identifying associations between germline mutations and clinicopathological characteristics. For example, it has been demonstrated that germline mutations in the genes involved in homologous recombination pathways, such as *BARD1*, *BRCA1*, *BRCA2*, *PALB2*, and *RAD51D*, are strongly associated with triple-negative breast cancer[3].

Most applications of NGS-based multigene panel testing focus only on small-sized variants containing single nucleotide variants (SNVs) and short insertions and deletions (INDELs). In addition, high-risk patients with hereditary breast and ovarian cancers harbor large germline rearrangements in *BRCA1/2*[4]. The effects of intermediate-sized INDELs (50–10,000 bp) on the pathogenicity of breast cancer remain uninvestigated due to technical difficulties in analyzing NGS data[5,6]. However, intermediate-sized INDELs are possibly involved in the pathology of breast cancer. Therefore, to clarify the clinical significance of intermediate-sized INDELs in breast cancer, we attempted to identify intermediate-sized INDELs in 54 cancer predisposition genes among 583 Han Chinese patients with breast cancer and identified a novel deletion–insertion in a translational variant of *PTEN* (also known as *PTENα* or *PTEN-Long*) in one patient.

Information regarding the study subjects and target-gene sequencing has been described in our previous study (Hata et al., submitted). In brief, 583 Han Chinese patients with breast cancer were recruited between December 2016 and September 2017 at the First Affiliated Hospital of Chongqing Medical University and Affiliated Cancer Hospital and Institute of Guangzhou Medical University. All patients provided informed consent for participation in this study. The Ethics Committees of the First Affiliated Hospital of Chongqing Medical University, the Affiliated Cancer Hospital & Institute of Guangzhou Medical

Correspondence: Ke Zheng (kezheng81@163.com) or Ituro Inoue (itinoue@nig.ac.jp) or Hua You (youhua307@163.com)
[1]Human Genetics Laboratory, National Institute of Genetics, Mishima, Japan
[2]The Department of Clinical Laboratory, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China
Full list of author information is available at the end of the article.

University, and the National Institute of Genetics approved the study protocols. The patients' mean age at diagnosis was 49.1 (standard deviation: 9.2) years.

Fifty-four cancer predisposition genes were selected based on previous studies of multigene panel testing for hereditary breast and/or ovarian cancer (Table S1). Target sequencing of these genes was performed using the pre-capture pooling method described in previous studies by using DNA samples isolated from peripheral blood[7,8]. The libraries were sequenced on an Illumina HiSeq 2500 platform operating in rapid-run mode using a 2 × 100-bp paired-end protocol (Illumina, San Diego, CA, USA).

NGS data processing and variant calling were performed using BWA[9] and GATK[10,11]. Functional annotation was implemented using ANNOVAR[12]. The estimation of variant frequencies in general populations was based on publicly available databases provided by ExAC[13]. Nonsense and splice-site SNVs and frameshifting INDELs were considered pathogenic. The variants with previously established pathogenic or benign effects were explored based on ClinVar[14]. We attempted to detect intermediate-to large-sized INDELs from mapped paired-end sequencing reads via bioinformatics analysis using Manta[15].

The average depth for the target regions was 117.6, and the mean proportion of the targeted regions covered by at least 20 reads was 99.3%, supporting confident variant detection. We identified 78 pathogenic mutations (43 SNVs and 35 short INDELs) in 21 genes containing *BRCA1/2* (Hata et al., submitted). However, pathogenic SNVs and short INDELs were not detected in 85.8% (500/583) of the patients.

Using Manta, we identified two intermediate-sized INDELs from patients without pathogenic SNVs or short INDELs. One was an 89-bp heterozygous deletion present in intron 14 of *APC* (NM_000038.5:c.1743 + 15_1743 + 103del89). The patient with this mutation was 53 years old and diagnosed with triple-negative breast cancer because estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor type 2 (HER2) were negative in immunohistochemistry (IHC) and fluorescence in situ hybridization. Adenomatosis polyposis coli (APC) is a tumor suppressor protein that acts as an antagonist of the Wnt signaling pathway. A lack of this gene causes familial adenomatous polyposis (OMIM175100). The identified intermediate-sized deletion in the intronic region of *APC* might change the splicing behavior of the gene. However, this deletion is registered as "likely benign" in ClinVar. The allele frequency of this deletion is 0.9% in the general East Asian population from the ExAC project. Based on these findings, the significance of this deletion for the pathogenicity does not seem high.

The other INDEL was a combination of a 47-bp deletion and a 68-bp insertion in *PTEN* (Fig. 1a) [NM_001304717.5 (PTEN_v001):c.8_54delinsAGTAATGTTAGCGGTTAGG CGTACGGCCAGGGCTATTGGTTGAATGAGTAGGC TGATGGTTTCGATAG (p. Arg3_Pro18delinsGlnTer); (Fig. 1a–d)]. Transcript NM_00130417 is a translational variant of *PTEN* (hereafter referred to as *PTENα*). The patient was heterozygous for this deletion–insertion. Although the result from Manta suggested the presence of deletion–insertion at *PTENα*, we could not exclude the possibility that the detected deletion–insertion was a false positive finding. Furthermore, the results from Manta did not provide any information about the possible source of the inserted sequence. To clarify these issues, we conducted additional bioinformatics analyses as described below. In addition, we performed Sanger sequencing to verify breakpoints of the deletion–insertion.

We attempted to determine the junctions of the intermediate-sized INDELs via in-house bioinformatics analysis that leverages split-reads of paired-end sequencing. First, we extracted soft-clipped reads containing a part of the unmatched sequence with the reference genome. Second, we divided the soft-clipped reads into unmatched and matched sequences by using an in-house Perl script. Third, we aligned these two types of read sequences with the reference genome by using BWA[9]. Finally, we searched the genomic positions where these reads were mapped and successfully determined the deletion junctions at *PTENα* from the realignment of the matched sequences of the soft-clipped reads (Fig. 1b, c). The presence of the unmatched sequences of the soft-clipped reads at the deletion junctions indicated that a DNA fragment derived from another region was inserted into the *PTENα* deletion site (Fig. 1d).

Next, we investigated the origin of the inserted DNA fragment (Fig. 2) by assembling the unmatched sequences of the soft-clipped reads at the deletion site of *PTENα* to determine a plausible sequence of the inserted DNA fragment. We then searched for the sequence against the human genome by using BLAT[16]. As a result, the inserted sequence matched two candidate regions: (i) a reverse complement of a region [chr1:569503–569570 (hg19)] within the nuclear mitochondrial sequence (chr1:564465–570304)[17] and (ii) a reverse complement of a part of the mitochondrial genome (chrM:8955–9022). Because the combinations of either of the two candidate-inserted segments with the PTENα sequence based on the human reference genome could not accurately account for the observed deletion–insertion, it resulted in one unresolved mismatch (G allele). Based on these results, we developed two hypotheses about the structure of the deletion–insertion. Hypothesis 1 assumed a 46-bp deletion and 67-bp insertion, in which the G allele originated from an alteration at the breakpoint of *PTENα*, whereas hypothesis 2 assumed a 47-bp deletion and 68-bp insertion, in which the G allele originated from an alteration in either of the two candidate insertions. Therefore, we

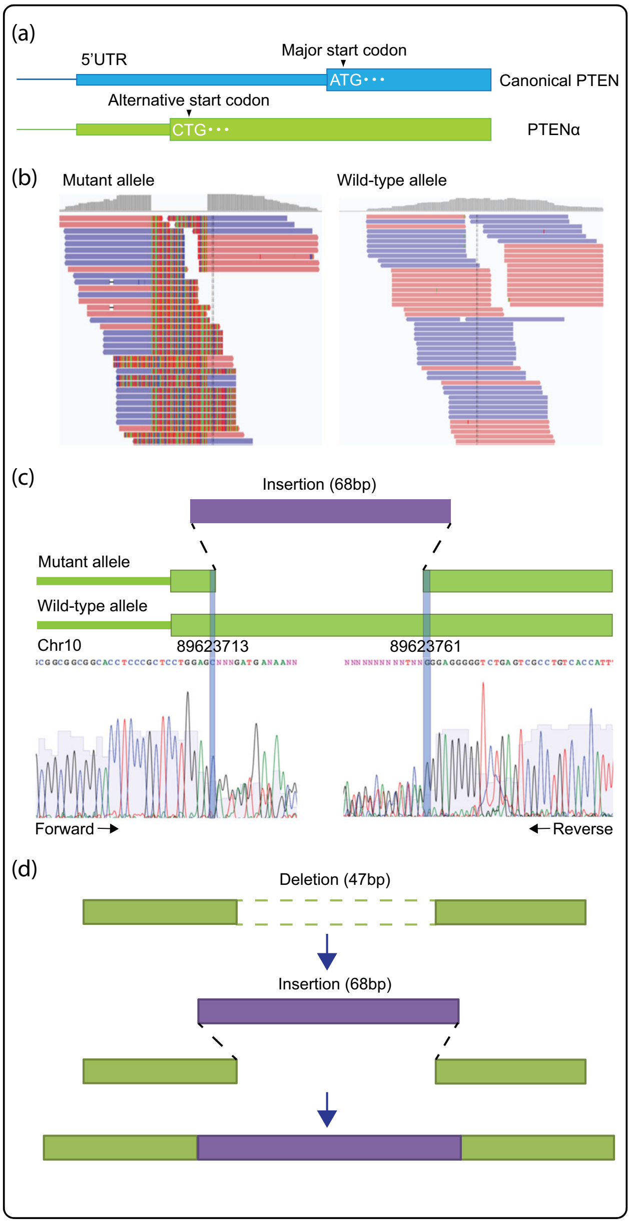

Fig. 1 Overview of the novel intermediate-sized INDEL in *PTENα*. **a** Differences in the structure between canonical *PTEN* (top) and *PTENα* (bottom). *PTENα* has an alternative start codon (CTG). **b** Detection of deletion junctions in *PTENα*. The alignment result of soft-clipped reads derived from the mutant allele (i.e., deletion–insertion) is shown. For comparison, the alignment result of reads that were not soft-clipped from the wild-type allele is also shown. The sequences color-coded by light red and blue indicate the sequences matched with that of the human reference genome. Highlighted bases indicate the sequences that are mismatched with the reference genome (i.e., inserted sequences). **c** A schematic representation with plausible junctions of the deletion–insertion in *PTENα*. The purple bar indicates an inserted sequence (68 bp), whereas the light green bars indicate *PTENα* sequences. Sanger sequencing confirmed the breakpoints of the deletion–insertion. Sanger sequencing using forward (left) and reverse (right) primers revealed aberrant electropherograms after the breakpoints of the deletion–insertion because the fluorescent signals from mutant and wild-type alleles were mixed. The breakpoints of the deletion–insertion are depicted as blue vertical lines. Sanger sequencing was performed using the BigDye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific, Waltham, MA, USA) on the ABI 3130x Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). The oligonucleotide sequences of the PCR primers are shown in Table S2. **d** A plausible mechanism of deletion and insertion at the same position. Light green and purple sequences indicate the reference and inserted sequences, respectively. Double-strand DNA breaks may result in a 47-bp deletion accompanied by a 68-bp insertion

scrutinized common genetic polymorphisms deposited in dbSNP. As a consequence, hypothesis 1 was rejected because there was no genetic polymorphism rendering the G allele at the breakpoint of *PTENα* (Fig. 2b). When considering hypothesis 2, the two candidate insertions were identical; however, a SNP located in the nuclear mitochondrial sequence on chromosome 1 (rs1198320487: NC_000001.10:g.569503T>C) differentiated the sequences (Fig. 2c). Finally, we determined that the G allele originated from the alternative allele of rs1198320487 in the nuclear mitochondrial sequence on chromosome 1.

The patient with the deletion–insertion in *PTENα* was diagnosed at 42 years of age, which is earlier than the average age of diagnosis in this study. IHC of ER and PR were negative, and we could not retrieve the results of HER2 from the patient's clinical charts. The frequency of the deletion–insertion of *PTENα* was not observed in any of the ExAC and other publicly available populations. The deletion–insertion was also not registered in either dbSNP or ClinVar, indicating that this mutation was a novel germline mutation. The identified deletion–insertion on exon 1 of *PTENα* was predicted to create a stop codon at the fourth amino acid of the PTENα protein.

PTEN is a tumor suppressor gene. *PTEN* mutations are commonly found in patients with inherited cancer syndromes, such as Cowden syndrome (OMIM158350). PTENα is a translational variant of *PTEN* and has an additional 173 amino acids at the N-terminus, labeled as alternatively translated region (ATR) (Fig. 1a). PTENα prevents cancer growth by antagonizing phosphoinositide-3 kinase signaling as well as canonical *PTEN*. More importantly, ATR contains a protein-binding domain and a cleavage site. These regions allow *PTENα* to bind to the cell membrane and be released into the extracellular space. Because ATR contains sequences that have homology with known cell-permeable peptides, *PTENα* enters into and acts in neighboring cells. Furthermore, PTENα without the cleavage site could not suppress tumor cell growth compared with normal PTENα in vivo[18]. From these results, we assumed that this novel protein-truncating mutation in *PTENα* could lead to the development of breast cancer due

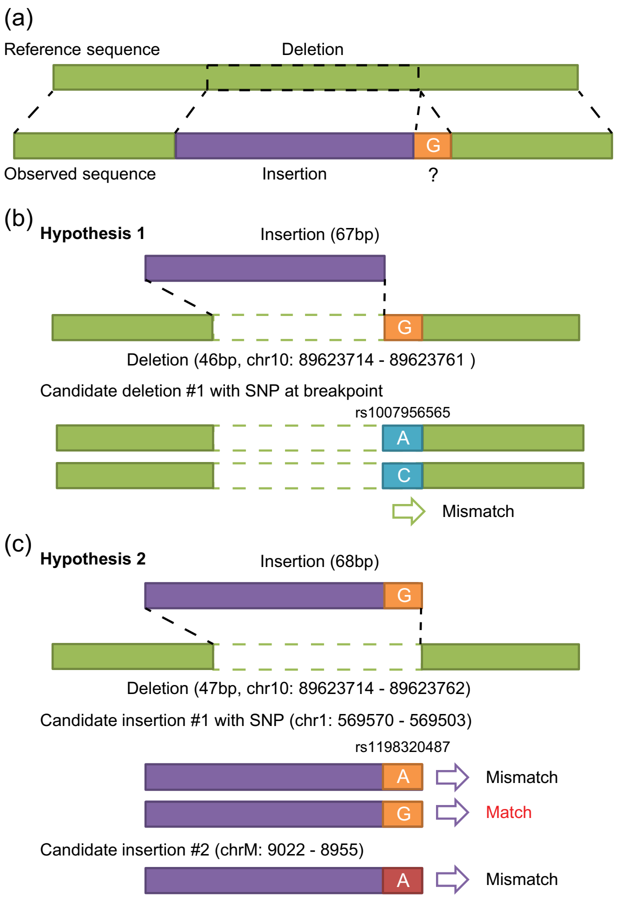
(a)
Reference sequence
Deletion
Observed sequence
Insertion
G
?
(b)
Hypothesis 1
Insertion (67bp)
G
Deletion (46bp, chr10: 89623714 - 89623761)
Candidate deletion #1 with SNP at breakpoint
rs1007956565
A
C
Mismatch
(c)
Hypothesis 2
Insertion (68bp)
G
Deletion (47bp, chr10: 89623714 - 89623762)
Candidate insertion #1 with SNP (chr1: 569570 - 569503)
rs1198320487
A
Mismatch
G
Match
Candidate insertion #2 (chrM: 9022 - 8955)
A
Mismatch

Fig. 2 Inference on the origin of the inserted sequence by the two hypotheses. a Differences in the structure between the reference (top) and observed (bottom) sequences. Light green and purple sequences indicate the reference and inserted sequences, respectively. The unresolved mismatch (G allele) around the breakpoint of *PTENα* is highlighted by an orange box. **b** Hypothesis 1 for the origin of the unresolved G allele. Hypothesis 1 assumes a 46-bp deletion and 67-bp insertion, in which the G allele arises from an alteration in the *PTENα* sequence by an SNP (rs1007956565, A/C). There are two possible sequences by the SNP (rs1007956565) at the breakpoint of *PTENα*; however, these two sequences cannot account for the unresolved mismatch (G allele). **c** Hypothesis 2 for the origin of the unresolved G allele. Hypothesis 2 assumes a 47-bp deletion and 68-bp insertion, in which the G allele arises from an alteration within the inserted sequence (purple). The two candidate sources of the inserted sequence are as follows: (i) a reverse complement of a region (chr1:569503–569570) within a nuclear mitochondrial sequence (chr1:564465–570304) and (ii) a reverse complement of a part of the mitochondrial genome (chrM: 8955–9022). These two candidate regions have identical sequences; however, there is a SNP (rs1198320487, A/G on the reverse strand) in the nuclear mitochondrial sequence on chromosome 1. One of the two possible sequences by the SNP (rs1198320487) can account for the unresolved mismatch (G allele). As a result, the source of the inserted sequence is likely to be the reverse complement of a region (chr1:569503–569570) within the nuclear mitochondrial sequence (chr1:564465–570304) with the alternative G allele at the SNP rs1198320487 site

to the lack of a tumor-suppressive function attributed to the PTENα protein. However, the functional significance of this deletion–insertion on canonical *PTEN* was unknown because the deletion–insertion was located on the 5′ untranslated region of canonical *PTEN* (Fig. 1a). Two pathogenic germline mutations on the N-terminal residues of *PTENα* were identified in a Chinese cohort of patients with autism spectrum disorder, although a definitive association between *PTENα* and neurodevelopment remains unknown[19].

Based on telephonic interviews and clinical charts, the patient with the deletion–insertion in *PTENα* did not report any family history of breast or ovarian cancer. Although we could not obtain DNA samples from her family members, there is a possibility that the deletion–insertion was inherited through her father's lineage. Other plausible explanations are that the mutation occurred de novo or arose at a very early stage of her development. Further examination of the genotypes of the mutation among her family members together with a review of her status is needed to assess the clinical significance of this novel deletion–insertion.

In conclusion, we identified a novel intermediate-sized deletion–insertion in *PTENα*, which can be a disease risk factor for breast cancer. This deletion–insertion may not be detected by general pipelines targeting SNVs and short INDELs in multigene panel testing. The breakpoint of the deletion and the possible source of the inserted fragment were determined by in-depth analyses. Therefore, our results suggest that patient-specific risk factors can be detected via detailed bioinformatics analyses.

HGV Database

The relevant data from this Data Report are hosted at the Human Genome Variation Database at https://doi.org/10.6084/m9.figshare.hgv.2786, https://doi.org/10.6084/m9.figshare.hgv.2789.

Acknowledgements

We are grateful to the participants in this study. The authors would like to thank Enago (www.enago.jp) for the English language review. This project was mainly supported by grants from Guangzhou Medical University (the start-up funds) and National Institute of Genetics (NIG Collaborative Research Program). H.Y. is supported by the National Natural Science Foundation of China (81911530169, 81903088, 81850410547, 81670180, and 81711540047), the Beijing Nova Program of the Beijing Municipal Science and Technology Commission (Z171100001117091), and the Guangzhou key medical discipline construction project. K.Z. is supported by the National Natural Science Foundation of China (81202090). I.I. is supported by JSPS and NSFC under the Japan-China Scientific Cooperation Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

Author details

[1]Human Genetics Laboratory, National Institute of Genetics, Mishima, Japan. [2]The Department of Clinical Laboratory, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China. [3]Chongqing University Cancer Hospital & Chongqing Cancer Institute & Chongqing Cancer Hospital, Chongqing, China. [4]School of Stomatology and Medicine, Foshan University, Foshan, Guangdong, China. [5]Affiliated Cancer Hospital & Institute of Guangzhou Medical University, Guangzhou, China. [6]Department of Endocrine and Breast Surgery, The First Affiliated Hospital of Chongqing Medical

University, Chongqing, China. [7]Affiliated Hospital of Academy of Military Medical Sciences, Beijing, China

Conflict of interest

The authors declare that they have no conflict of interest.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information is available for this paper at https://doi.org/10.1038/s41439-019-0080-8.

Received: 4 June 2019 Revised: 9 September 2019 Accepted: 14 September 2019.
Published online: 29 October 2019

References

1. DeSantis, C. E., Ma, J., Goding Sauer, A., Newman, L. A. & Jemal, A. Breast cancer statistics, 2017, racial disparity in mortality by state. *CA Cancer J. Clin.* **67**, 439–448 (2017).
2. Ellisen, L. W. & Haber, D. A. Hereditary breast cancer. *Annu Rev. Med.* **49**, 425–436 (1998).
3. Shimelis, H. et al. Triple-negative breast cancer risk genes identified by multigene hereditary cancer panel testing. *J. Natl. Cancer Inst.* https://doi.org/10.1093/jnci/djy106 (2018).
4. Judkins, T. et al. Clinical significance of large rearrangements in BRCA1 and BRCA2. *Cancer* **118**, 5210–5216 (2012).
5. Claudia, M. B., Carvalho & Lupski, JamesR. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
6. Shigemizu, D. et al. IMSindel: an accurate intermediate-size indel detection tool incorporating de novo assembly and gapped global-local alignment with split read analysis. *Sci. Rep.* **8**, 5608 (2018).
7. Ahmadloo, S. et al. Rapid and cost-effective high-throughput sequencing for identification of germline mutations of BRCA1 and BRCA2. *J. Hum. Genet.* **62**, 561–567 (2017).
8. Suda, K. et al. Clonal expansion and diversification of cancer-associated mutations in endometriosis and normal endometrium. *Cell Rep.* **24**, 1777–1789 (2018).
9. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
10. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
11. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
12. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
13. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
14. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**(Database issue), D862–D868 (2016).
15. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
16. Kent, W. J. BLAT-the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
17. Simone, D., Calabrese, F. M., Lang, M., Gasparre, G. & Attimonelli, M. The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. *BMC Genomics*. https://doi.org/10.1186/1471-2164-12-517 (2011).
18. Hopkins, B. D. et al. A secreted PTEN phosphatase that enters cells to alter signaling and survival. *Science* **341**, 399–402 (2013).
19. Zhou, W. Z. et al. Targeted resequencing of 358 candidate genes for autism spectrum disorder in a Chinese cohort reveals diagnostic potential and genotype-phenotype correlations. *Hum. Mutat.* **40**, 801–815 (2019).